6W-02

レビュー投稿における不快表現の通知機能システム

鈴木会子　浦川紗貴　松井美結　中野美由紀
津田塾大学学芸学部情報科学科

1.　はじめに

近年，店や施設を訪れた感想やコメントをレビューとして残す人が増え，また施設を選ぶ際にレビューや口コミをみて判断する人も多い．SNS などを通じ，自由にコメントできる結果，過激あるいは高圧的な表現など閲覧者が不快に感じる投稿も散見される[1]．

本システムはレビュー・コメント投稿時に，閲覧者が不快になる可能性が高い表現を使用する可能性を通知し，投稿者が文を見直す機会をつくることを目的としている．楽天トラベルのレビュー記事を基に不快と感じられるレビューの抽出を人手で行った．不快なレビュー記事から特徴語を抽出し，その特徴語をベースに不快と感じられる語の辞書を word2vec を用いて作成し[2,3]，その評価を行った．投稿時に不快と感じられる語が用いられた場合，不快となる可能性を通知し，投稿者に注意を促すことに利用する．

2.　不快表現通知機能システム

2.1　システムの実装方式

実装には Python のウェブアプリケーションフレームワーク Django を使用した．

2.2　アプリケーション詳細

タイトルとレビュー本文のどちらかあるいは両方に不快な表現が含まれていた場合，「送信」ボタンを押すと入力フォームの下にアラート文が表示される．そしてアラート文に従い不快な表現を改めてから，あるいは不快な表現を含んだまま再度「送信」ボタンを押すことで，トップ画面へと遷移しレビューが投稿される．

3.　不快語検出アルゴリズム

楽天トラベルの 2004～2019 年総評価星 1 のレビューデータからランダムに選んだ 200 件のレビュー記事を人手で不快であるかどうか判定・タグ付した．このデータを用い，不快であると判定したレビューの特徴語を不快語として抽出する．レビューごとに全単語数における不快語数を計算することで不快レビューの検出を行う．

3.1　アルゴリズムに用いるデータ

提案手法で用いるデータは，楽天技術研究所が公開する楽天データ[4]のうち，2004 年から 2019 年までの楽天トラベルのレビューデータ（レビュー数 6560173 件）である．本研究では，不快と感じられるレビューの抽出を行うために総評価星 1 のレビューデータ（109248 件）を使用する．図 1 に 2004 年から 2019 年までのレビューデータの星の分布を示す．

図 1．2004～2019 年のレビューデータ星分布

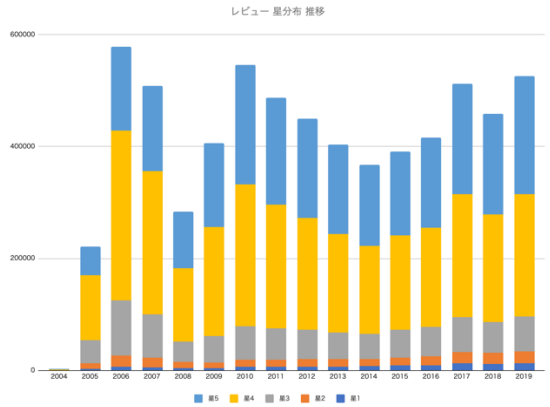

図 1 より年によって星分布に大きな差異は見られない．2014 年のレビューから，星 1 から星 5 までそれぞれ 100 件ずつランダムに選び人手で快・不快のタグ付を行い，不快と判断されたレビューは星 1 が 8 割であった．よって，以降不快レビューの収集には総評価が星 1（図 1 内濃青部分）のレビューを用いてタグ付を行った．

3.2　アルゴリズムの流れ

3.2.1　不快と判定されたレビューからの特徴語の抽出

a.　人手で評価した不快レビューと非不快レビューで TF-IDF を実施し，不快語辞書のベースを作成した．この際，TF-IDF の閾値は 0.1 で判定を行い，表 1 に示す 10 語が不快レビューの特徴語として抽出できた．

b.　表 1 の 1 0 語を基に Word2Vec を利用して，類義語を得た．類似度が 0.6 以上の単語を不快語の候補として評価を行った．類似度 0.6 以上で 567 語，0.7 以上で 163 語，0.8 以上で 41 語，0.9 以上で 17 語抽出することができた．表 2 にその結果の一部を示す．

A notification system of unpleasant expressions when posting a review
Aiko Suzuki, Saki Urakawa, Miyu Matsui, Miyuki Nakano・Tsuda University

表１．不快レビューの特徴語

お客様	0.18972652
こと	0.16601071
ため	0.11857908
ない	0.16601071
ビール	0.16665875
フロント	0.23715815
ホテル	0.16601071
夕食	0.11857908
最低	0.14229489
部屋	0.45060049

表２．不快１０語から得られた類義語（一部）

お客様	ゲスト	0.811469018
お客様	客	0.752257109
お客様	客人	0.726819634
お客様	彼ら	0.650649428
お客様	お客	0.64263469
お客様	顧客	0.629567862
こと	事	0.978707492
こと	コト	0.754916549
こと	わけ	0.61645329
こと	ハプニング	0.611406147
ため	為	0.983383179
ため	ので	0.650216162
ため	せい	0.602443993
ない	なかっ	0.801666439
ない	無い	0.792243421
ない	なく	0.675818384

3.2.2 不快レビューの検知

ユーザーがレビューを投稿する時，そのレビューの全単語数と 3.2.1 で作成した不快語辞書に含まれている単語数を数える．全単語数における不快語数の割合を計算し，その割合が閾値以上のものを不快レビューと注意を促す.

4. 評価

評価データには，総評価星ごとの不快レビュー分布調査でタグ付したデータを用いた．3.2.2 において閾値を 1.0,2.0,3.0,4.0,5.0%とし，本手法で検出したレビュー数の割合を以下に示す.

表３．評価結果

Word2vec 類似度 0.6 以上

閾値	不快レビューの中からの検出割合	快レビューからの検出割合
1.00%	100.00%	81.80%
2.00%	100.00%	75.06%
3.00%	90.00%	64.94%
4.00%	70.00%	51.01%
5.00%	50.00%	33.26%

Word2vec 類似度 0.7 以上

閾値	不快レビューの中からの検出割合	快レビューからの検出割合
1.00%	100.00%	74.61%
2.00%	80.00%	62.47%
3.00%	60.00%	47.87%
4.00%	40.00%	32.36%
5.00%	30.00%	20.45%

Word2vec 類似度 0.8 以上

閾値	不快レビューの中からの検出割合	快レビューからの検出割合
1.00%	90.00%	62.25%
2.00%	70.00%	44.04%
3.00%	60.00%	29.21%
4.00%	30.00%	18.20%
5.00%	30.00%	10.56%

Word2vec 類似度 0.9 以上

閾値	不快レビューの中からの検出割合	快レビューからの検出割合
1.00%	90.00%	58.43%
2.00%	60.00%	40.22%
3.00%	50.00%	23.82%
4.00%	30.00%	13.93%
5.00%	30.00%	8.31%

あるレビューの全単語数と 3.2.1 で作成した不快語辞書に含まれている単語数を数える．全単語数における不快語数の割合を計算し，その割合が閾値以上のものを不快レビューとして扱う.

4. 考察

不快レビューの可能性を示唆する場合，快レビューにおいても高い頻度で不快の可能性がある単語が含まれていることが，表３から分かった．例えば，「こと」などの言葉は快・不快を問わず利用される可能性がある．一方で，類似度をあげると，快レビューにおいて不快レビューの注意を出す可能性が下がることも確認できた．本通知機能は注意を促すために利用するものであり，書き手の判断にゆだねる部分も必要と思われる.

5. まとめ

本研究では，不快なレビューの特徴語を用いて投稿者に表現の改善を促し，レビューや口コミを通じた閲覧者の不快感を減らすシステムの提案・開発を行った．今後はフレーズなども考慮し，不快部位を詳細に提示できるシステムに改善していきたい.

参考文献

[1]三宅 剛史，松本 和幸，吉田 稔，北 研二：分散表現を用いた有害表現判別に基づく炎上予測，人工知能学会 インタラクティブ 情報アクセスと可視化マイニング研究会(第 15 回)，2017-03-03

[2]村上奈緒，尼岡利崇：Twitter 上で任意の検索語句に対するネガポジ度を判定し可視化するアプリケーションの開発と研究，エンタテインメントコンピューティングシンポジウム 2014 論文集，2014-09-12

[3]藤平 啓汰：感情辞書を用いた Web テキストの多言語感情抽出，第 82 回全国大会講演論文集，2020-02-20

[4]楽天技術研究所：楽天データ公開 https://rit.rakuten.com/data_release_ja/

謝辞

本研究は JSPS 科研費 18K11318 の助成を受けたものです.